



Full length article

Multimodal sentiment analysis with unimodal label generation and modality decomposition

Linan Zhu^a, Hongyan Zhao^a, Zhechao Zhu^a, Chenwei Zhang^b, Xiangjie Kong^{a,*}

^a College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China

^b School of Faculty of Education, University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:

Multimodal sentiment analysis

Unimodal label generation

Modality decomposition

ABSTRACT

Multimodal sentiment analysis aims to combine information from different modalities to enhance the understanding of emotions and achieve accurate prediction. However, existing methods face issues of information redundancy and modality heterogeneity during the fusion process, and common multimodal sentiment analysis datasets lack unimodal labels. To address these issues, this paper proposes a multimodal sentiment analysis approach based on unimodal label generation and modality decomposition (ULMD). This method employs a multi-task learning framework, dividing the multimodal sentiment analysis task into a multimodal task and three unimodal tasks. Additionally, a modality representation separator is introduced to decompose modality representations into modality-invariant representations and modality-specific representations. This approach explores the fusion between modalities and generates unimodal labels to enhance the performance of the multimodal sentiment analysis task. Extensive experiments on two public benchmark datasets demonstrate the effectiveness of this method.

1. Introduction

With the advent of the big data era and the continuous advancement of deep learning technologies, numerous deep learning methods have achieved significant development. For instance, UIU-Net has notably improved the detection of small objects in infrared images [1], DC-Net effectively integrates the intrinsic characteristics of hyperspectral and multispectral images [2], and CCR-Net consolidates different modal features extracted through CNNs in a more compact manner [3]. In line with this trend, sentiment analysis tasks have also expanded beyond textual data to encompass multimodal data sources, including text, image, and audio modalities. Multimodal sentiment analysis (MSA) leverages information from image and audio modalities to assist text-based predictions. The text modality provides the semantic meaning of speech, the image modality extracts facial features of the speaker (such as facial expressions and gestures), and the audio modality reflects the emphasis and intensity of speech (via, e.g., pitch and volume). Multimodal sentiment analysis systems that incorporate multiple modalities consistently outperform the best unimodal sentiment analysis systems [4]. By analyzing images, audio, and text expressions, it becomes possible to better understand human emotional communication and pave the way for more human-like artificial intelligence [5,6]. Recent methods, such as RustQNet, which analyzes multimodal remote sensing images for the quantitative inversion of the wheat stripe rust

disease index [7], and HighDAN, which utilizes the multimodal remote sensing dataset (C2Seg) to enhance model generalization and segmentation performance in cross-city environments [8], further validate the effectiveness of multimodal techniques in enhancing the understanding capabilities of various applications.

Multimodal fusion is a crucial step in enabling artificial intelligence to understand human emotions. Early methods merged features immediately after extraction, which, although simple and direct, failed to adequately integrate useful information from different modalities and introduced a significant amount of redundant data, leading to a decline in the model's learning efficiency [9,10]. To address this issue, late fusion emerged, employing more complex model structures that process each modality's data through their respective independent models before merging the results into a final vector [11]. Hybrid fusion combines the advantages of both early and late fusion, allowing features to be integrated at multiple stages within the model, but it has a higher complexity and still requires further optimization [12]. Despite the advancements made in multimodal fusion, they still face issues of information redundancy and significant differences in the representation of information between different modalities. Additionally, each instance in multimodal datasets typically contains only one multimodal sentiment label, and the lack of unimodal labels limits the effective learning of modality-specific information.

* Corresponding author.

E-mail address: xjkong@ieee.org (X. Kong).

<https://doi.org/10.1016/j.inffus.2024.102787>

Received 2 September 2024; Received in revised form 13 October 2024; Accepted 5 November 2024

Available online 20 November 2024

1566-2535/© 2024 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

To address these issues, this paper proposes a multimodal sentiment analysis method based on unimodal label generation and modality decomposition (ULMD). This approach uses a multi-task learning framework to divide the entire model into one multimodal and three unimodal tasks [13,14], as shown in Fig. 1. First, after feature extraction, different encoders are used to encode different modalities for use in both types of tasks. Second, in the multimodal task, text is used as the core, and through a multilayer perceptron module, it is fused with visual and audio modalities to form sentiment polarity vectors and sentiment intensity vectors, which are then multiplied to obtain a multimodal sentiment representation. For unimodal tasks, three modality-specific multilayer perceptrons are used to obtain sentiment representations for each modality. Then, combining the multimodal feature representations, multimodal sentiment labels and unimodal feature representations, unimodal supervision labels for each modality are generated. Finally, joint training of the multimodal task and the three unimodal tasks improves the model's performance.

However, a new problem emerged during our implementation. Since information from different modalities is processed in a shared underlying network, the unique information of specific modalities might be diluted or lost, hampering the model's ability to fully leverage it. To solve this problem, we introduce a modality representation decomposer in ULMD. This decomposer splits each modality representation into modality-invariant and modality-specific representations. The modality-invariant representations are fed into the multimodal task, while the modality-specific representations are fed into the unimodal tasks. In this way, the model can simultaneously obtain extensive common information and in-depth individual information, resulting in better performance in sentiment analysis tasks.

We conducted experiments mainly on two benchmark datasets of MSA: CMU-MOSI and CMU-MOSEI. The experiments show that our model outperforms the current state-of-the-art models on most metrics. Furthermore, ablation experiments demonstrate the effectiveness of each module in our approach.

The main contributions of this paper can be summarized as follows:

- We proposed a modality representation separation module that decomposes modality representations into modality-invariant and modality-specific representations. This addressed the issue of unique modality information being diluted or lost during the fusion process, thereby maintaining each modality's consistency during multimodal task fusion and preserving its independence in unimodal tasks.
- We designed a fusion module centered on the text modality for multimodal tasks, using multilayer perceptrons to explore interactions between different modalities, which resulted in sentiment polarity and sentiment intensity representations. We also introduced an auxiliary task for generating unimodal sentiment labels in unimodal tasks, addressing the issue of lacking unimodal labels in commonly used datasets.
- We conducted extensive experiments on two benchmark datasets, demonstrating that our method outperforms the current state-of-the-art models.

2. Related work

In this section, we discuss related work in multimodal sentiment analysis and multimodal representation learning, highlighting the innovations of our approach.

2.1. Multimodal sentiment analysis

Multimodal sentiment analysis is a relatively new research field that aims to predict sentiment polarity by integrating multiple modalities (text, image, and audio). Early methods in multimodal sentiment analysis can be categorized into early fusion, late fusion, and tensor-based

fusion [15,16]. For early fusion, Morency et al. [17] first introduced a task for trimodal sentiment analysis. They extracted features from each modality, concatenated them, and used a trimodal HMM classifier to learn the hidden structure of the input signals. Perez-Rosas et al. [9] combined features collected from all multimodal streams into a single feature vector and used an SVM classifier to determine the sentiment orientation of the utterance. In an example of late fusion, Nojavanasghari et al. [18] trained a unimodal classifier for each modality and then averaged the confidence scores of each unimodal classifier to make the final prediction. To carry out tensor-based fusion, Zadeh et al. [19] proposed a Tensor Fusion Network (TFN) model, which applied a triple Cartesian product to the three output vectors from the embedding layer, fully integrating unimodal, bimodal and trimodal interactions. Liu et al. [20] introduced a Low-rank Multimodal Fusion (LMF) method that decomposed weights into low-rank factors for multimodal fusion, improving computational efficiency.

Researchers often focus on designing complex model frameworks to integrate multimodal information. However, after extracting features from each modality for fusion, they tend to overlook the original feature information, forgoing exploration of interactions and influences between unimodal and multimodal features. Unlike previous studies, our method employs multilayer perceptrons to explore such interactions, using the text modality as the core. It interacts with video and audio modalities to obtain sentiment polarity and intensity representations, which are then fused to achieve the final sentiment classification result.

2.2. Multimodal representation learning

Representation learning, which involves learning representations of data, can automatically extract and organize discriminative information from data [21]. Effective representation learning can transform data into more useful and abstract forms applicable to various downstream tasks. Baltrušaitis et al. [22] identified five core challenges in multimodal learning: alignment, translation, representation learning, fusion and co-learning. Since representation learning directly affects the other four challenges, it is considered fundamental. Hazarika et al. [23] proposed the MISA framework, which projects each modality into a modality-invariant subspace to reduce modality gaps and a modality-specific subspace to capture its characteristic features. Yang et al. [24] introduced a ConFEDE method that enhances multimodal information representation through contrastive representation learning and contrastive feature decomposition. In this work, we designed a modality representation separator that decomposes input modalities into modality-invariant and modality-specific representations, preserving both the common and unique information of each modality.

2.3. Unimodal label generation

Unimodal labels and multimodal labels represent the sentiment or category annotations for uni- and multimodal data, respectively. In multimodal sentiment analysis, unimodal labels can help the model better understand the characteristics of each modality, while multimodal labels provide a comprehensive perspective to more accurately capture sentiments or other information. Most existing multimodal sentiment analysis datasets only contain an overall multimodal sentiment label and lack modality-specific annotations, limiting the training of unimodal sentiment analysis. Therefore, unimodal label generation has recently been widely applied in MSA. Yu et al. [13] designed a weight self-adjustment strategy to guide subtasks to focus on samples with large differences between modality supervisions, balancing the learning progress between different subtasks. Hwang et al. [25] developed a SUGRM framework, which performs MSA using a self-supervised unimodal label generation strategy with recalibrated modality representations. M. Li et al. [26] demonstrated a fusion strategy for joint training of unimodal and multimodal (JTUM), combining a unimodal

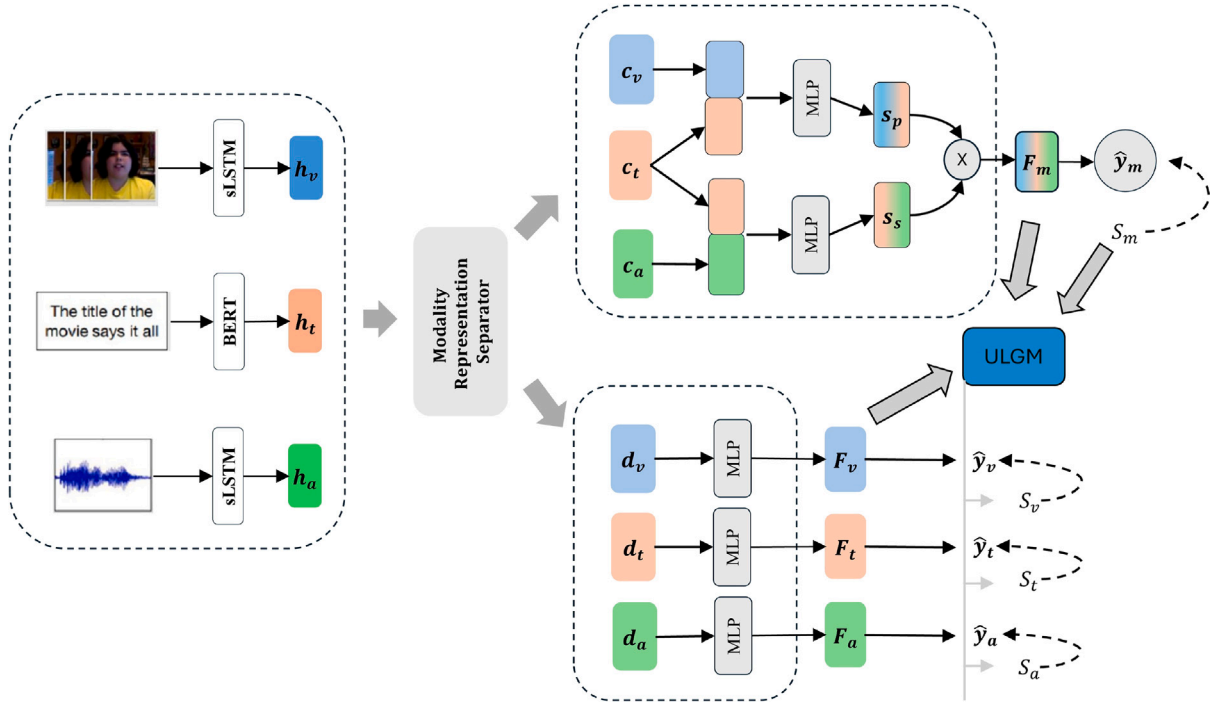


Fig. 1. The structure of the ULMD framework. h_u represents the original features, c_u represents the modality-invariant representations, and d_u represents the modality-specific representations, where $u \in \{t, a, v\}$. S_p and S_s are the sentiment polarity vector and sentiment intensity vector, respectively. \hat{y}_m , \hat{y}_t , \hat{y}_a and \hat{y}_v are the prediction outputs for the multimodal task and the three unimodal tasks. Finally, S_m is the human-annotated multimodal label. S_t , S_a and S_v are the unimodal supervision signals generated through a self-supervised strategy.

label generation module and a cross-modal Transformer. Z. Li et al. [27] proposed a multimodal sentiment analysis method based on multilevel relevance mining and self-supervised multi-task learning. Inspired by the work of Yu et al. [13], this paper designs a Unimodal Label Generation Module (ULGM) to generate unimodal label supervision information. In this process, the positive and negative centers of each modality are calculated, and cosine similarity is used to determine directionality, ensuring a proportional mapping relationship between modality representations and label supervision values.

3. Methodology

3.1. Task description

The task of multimodal sentiment analysis involves using multi-modal data, including text (X_t), audio (X_a) and image (X_v), to perform sentiment analysis on review content. Generally, this task can be considered either a regression task or a classification task. Therefore, the task is formulated as taking X_t , X_a and X_v as inputs and producing either a sentiment intensity result $y_m \in \mathbb{R}$ or a predicted classification label. The model framework for this method is shown in Fig. 1.

3.2. Modality representation module

3.2.1. Feature representation

In the feature representation layer, we use pre-trained models and open-source toolkits to extract features from text, image, and audio data. For the image and audio modalities, we use the open-source toolkits OpenFace and COVAREP to extract initial vector features from the raw data: specifically, the initial vector features for image modality $X_v \in \mathbb{R}^{l_v \times d_v}$ and audio modality $X_a \in \mathbb{R}^{l_a \times d_a}$, where $l \times d$ represent the feature dimensions. Since the sequence of image and audio segments affects the semantic expression and has temporal relevance, we use a single-layer Long Short-Term Memory network (sLSTM) to embed the

contextual relationships, obtaining image modality embeddings h_v and audio modality embeddings h_a . The formulas are as follows:

$$h_v = \text{sLSTM}(X_v, \theta_v) \quad (1)$$

$$h_a = \text{sLSTM}(X_a, \theta_a) \quad (2)$$

where θ_v and θ_a represent the network parameters for the image and audio modalities, respectively.

For the text modality, we use a BERT model to encode it, resulting in the representation of the entire sentence h_t . The formula is as follows:

$$h_t = \text{BERT}(X_t, \theta_t) \quad (3)$$

where X_t represents the raw text, and θ_t represents the network parameters for the text modality.

3.2.2. Modality decomposition

After extracting features from the three modalities, we input them into the modality representation decomposition module. The process of modality separation is shown in Fig. 2. This module comprises two components: modality-invariant encoders and modality-specific encoders. The modality-invariant encoders map the feature vectors of each modality into modality-invariant representations. These representations capture common features across different modalities, reducing modality differences and highlighting the consistency between modalities. This helps to make the fusion process smoother and more effective when integrating the information of these modalities. The modality-specific encoder maps the feature vectors into modality-specific representations, focusing on the unique characteristics or differential information of each modality. Different modalities may contain important information that other modalities do not possess. For instance, the visual modality may include information about facial expressions, while the audio modality may contain information about tone and emotion. The formulas are as follows:

$$c_u = E_c(h_u) \quad (4)$$

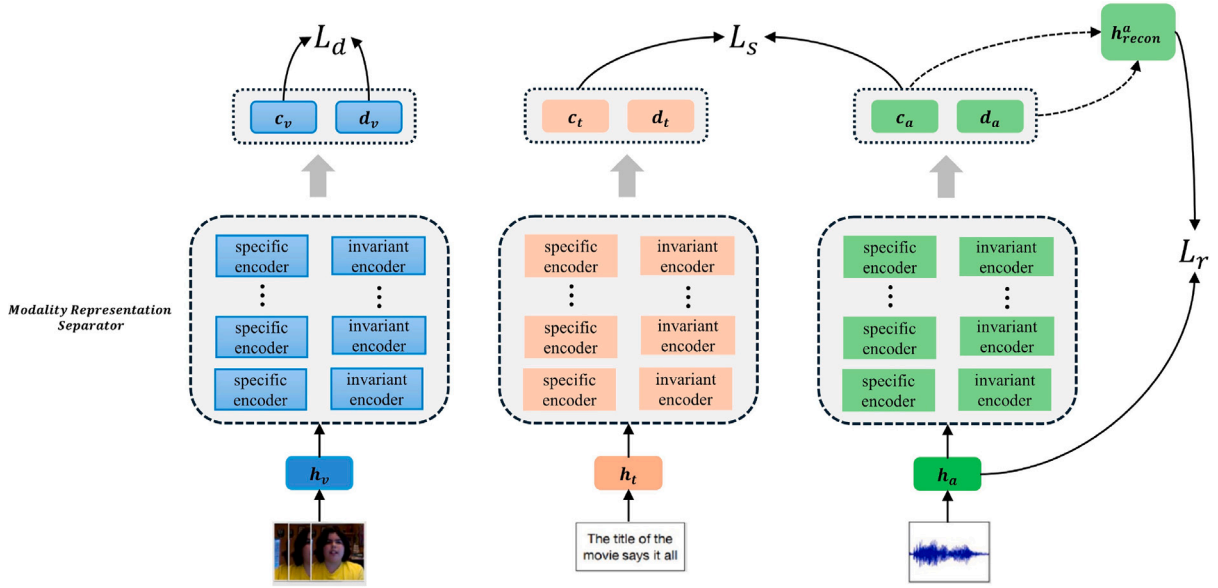


Fig. 2. The modality separator divides the input features into modality-invariant representations and modality-specific representations. h_{recon}^a represents the reconstructed audio features, L_s denotes the similarity loss, L_d denotes the diversity loss, and L_r denotes the reconstruction loss.

$$d_u = E_d(h_u) \quad (5)$$

where $u \in \{t, a, v\}$, h_u represents the original features, c_u represents the modality-invariant representation, and d_u represents the modality-specific representation.

3.3. Multimodal task

After mapping the three modalities to their respective representations, we use the modality-invariant representations for the multimodal task. Previous studies have demonstrated that in multimodal sentiment analysis tasks, the text modality often plays the most crucial role, providing the most accurate and comprehensive information [28, 29]. Therefore, in this work, the multimodal task centers on the text modality and fully integrates it with the other two modalities through interaction.

Specifically, we use a multilayer perceptron (MLP) to facilitate the interactions between modalities, generating two representation vectors: the sentiment polarity vector (S_p) and the sentiment intensity vector (S_s). The image modality, which contains rich information about facial expressions and head poses, can supplement the text modality in determining non-conventional emotions (such as sarcasm and exaggeration). Through the interaction between the image and text modalities, we obtain S_p to determine the sentiment polarity (positive, neutral, or negative). Simultaneously, the audio modality, which includes features such as pitch and volume, helps in assessing the intensity of the emotion. Hence, through the interaction between the audio and text modalities, we obtain S_s to determine the emotion's intensity, which is crucial, in a 7-point scale classification. The formulas are as follows:

$$S_p = \text{MLP}_p(c_t; c_v, \theta_p) \quad (6)$$

$$S_s = \text{MLP}_s(c_t; c_a, \theta_s) \quad (7)$$

where MLP_p and MLP_s each consist of a linear layer with different weight parameters, a ReLU activation function, and a normalization layer. θ_p and θ_s represent the weight parameters of MLP_p and MLP_s , respectively.

After obtaining the sentiment polarity vector S_p and the sentiment intensity vector S_s , we extract the sentiment direction from S_p and the sentiment intensity from S_s . We then use a simple multiplication to fuse

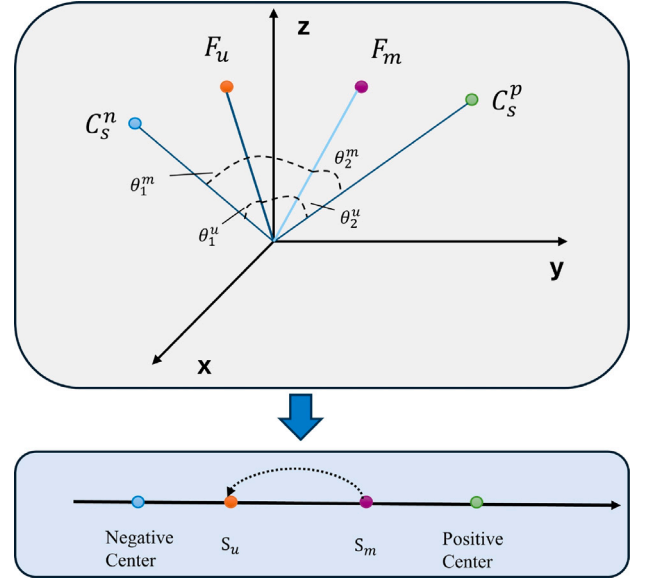


Fig. 3. Unimodal label generation example. θ_1^m represents $\arccos D_{mp}$, θ_2^m represents $\arccos D_{mp}$, θ_1^u represents $\arccos D_{mp}$, and θ_2^u represents $\arccos D_{up}$. The unimodal label S_u is obtained by adding an offset to the multimodal label S_m .

them into a composite vector F_m , which represents the final multimodal feature for prediction. The formulas are as follows:

$$F_m = \frac{S_p}{\|S_p\|_2} \times \|S_s\|_1 \quad (8)$$

$$\hat{y}_m = W^m F_m + b^m \quad (9)$$

3.4. Unimodal tasks and label generation task

We use the modality-specific representations for unimodal tasks. Different modality-specific MLP classifiers explore the modality-specific information contained in each modality and input it into a linear layer

to obtain the unimodal sentiment prediction \hat{y}_u . The formulas are as follows:

$$F_u = \text{MLP}_u(d_u, \theta_u) \quad (10)$$

$$\hat{y}_u = W^u F_u + b^u \quad (11)$$

where $u \in \{t, a, v\}$, θ_u represents the weight parameters of MLP_u , and W^u and b^u are the weight parameters to be trained.

The ULGM is introduced to carry out unimodal label supervision by effectively utilizing multimodal label annotations and unimodal representation information, thereby generating unimodal label annotations, as shown in Fig. 3. In this label generation process, two correlations are considered. First, the unimodal labels are highly correlated with the multimodal labels. Second, there is a proportional mapping between the modality representations and the modality label supervision.

Regarding the first correlation, in general, the sentiment polarity of unimodal and multimodal labels remains consistent. However, in the real world, due to expressions such as sarcasm and exaggeration, the sentiment judgment and generated labels of each unimodal modality may differ from the multimodal ones. For example, the phrase “You are amazing” would be positively labeled based solely on the text modality, but if paired with a frown and exaggerated facial expressions, the image modality might be labeled negative, and considering the tonal changes in the audio modality, the final dataset might be annotated as negative. Therefore, to identify samples where the unimodal labels differ from the multimodal labels, we calculate the positive and negative centers of each unimodal and multimodal feature representation:

$$C_s^p = \frac{\sum_{i=1}^N I(y_s(i) > 0) \cdot F_{si}}{\sum_{i=1}^N I(y_s(i) > 0)} \quad (12)$$

$$C_s^n = \frac{\sum_{i=1}^N I(y_s(i) < 0) \cdot F_{si}}{\sum_{i=1}^N I(y_s(i) < 0)} \quad (13)$$

where $s \in \{m, t, a, v\}$, N is the number of training samples, $I(\bullet)$ is an indicator function, and F_{si} is the modality representation of the i th sample of modality s .

Cosine similarity is used to calculate the similarity between modality representations and positive/negative centers in terms of direction. Cosine similarity measures the directional similarity between two vectors in vector space. In high-dimensional spaces, many distance metrics (such as Euclidean distance) become unreliable because data points are often very far apart. Cosine similarity is chosen to better capture the similarity between vectors in high-dimensional space. The formulas are as follows:

$$D_{sp} = \frac{F_{si} \cdot C_s^p}{\|F_{si}\| \|C_s^p\|} \quad (14)$$

$$D_{sn} = \frac{F_{si} \cdot C_s^n}{\|F_{si}\| \|C_s^n\|} \quad (15)$$

where $s \in \{m, t, a, v\}$, F_{si} is the modality representation of the i th sample of modality s , D_{sp} represents the similarity of modality s to the positive center in terms of direction, and D_{sn} represents the similarity of modality s to the negative center in terms of direction.

We define the relative distance value α , which represents the relative distance of the modality representation to the positive center C_s^p and negative center C_s^n . The formula is as follows:

$$\alpha = \frac{D_{sn} - D_{sp}}{\|F_{si}\| (\|C_s^p\| + \|C_s^n\|) + \epsilon} \quad (16)$$

where ϵ is used to prevent zero-value anomalies.

For the second correlation, there is a certain positive mapping between the modality representation and the modality supervision value. The relationship can be expressed as follows:

$$F_m(S_m) \propto F_u(S_u) \quad (17)$$

where (\cdot) represents some mapping function, $u \in \{t, a, v\}$, S_m and S_u are the supervision values of the multimodal and unimodal representations, respectively, and \propto indicates a direct proportional relationship.

Furthermore, in this work, by considering the offset difference, we represent the mapping function through ratio and difference. The formulas are as follows:

$$\alpha_m - S_m = \alpha_u - S_u \quad (18)$$

$$\frac{\alpha_m}{S_m} = \frac{\alpha_u}{S_u} \quad (19)$$

Thus, the calculation for the unimodal label supervision value S_u proceeds as follows:

$$S_u = \frac{\alpha_u - \alpha_m + S_m}{2} + \frac{\alpha_u \cdot S_m}{2\alpha_m} \quad (20)$$

$$S_u = S_m + \frac{\alpha_u - \alpha_m}{2} \cdot \frac{S_m + \alpha_m}{\alpha_m} \quad (21)$$

where $u \in \{t, a, v\}$.

Due to the instability of generating unimodal labels, a dynamic adjustment strategy is adopted to synthesize the newly generated labels with historical labels, giving less weight to later generated label supervision.

$$S_u^{(i)} = \begin{cases} S_m & i = 1 \\ \frac{1}{2} S_u^{(1)} + \frac{1}{2} S_u^{(2)} & i = 2 \\ \frac{i-1}{i+1} S_u^{(i-1)} + \frac{2}{i+1} S_u^i & i > 2 \end{cases} \quad (22)$$

where $u \in \{t, a, v\}$, S_u^i is the label supervision of modality u generated in the i th round, and $S_u^{(i)}$ is the final label supervision of modality u after the i th round.

Algorithm 1 ULMD

Input: Unimodal features h_u and multimodal labels S_m

Output: Prediction \hat{y}_m

while not done do

 Sampling a batch of data

for each sample i do

 Generate c_u , d_u , F_m and F_u based on h_u

 Compute positive and negative centers C_s^p and C_s^n , as in Eqs. (12) and (13)

 Compute the directional similarity between each modality representation and the positive and negative centers as in Eqs. (14) and (15)

 Update unimodal labels S_u based on S_m , F_u and F_m , as in Eq. (16) to (22)

 Generate prediction \hat{y}_m

 Compute losses L_s , L_d and L_r based on c_u , d_u and h_u

 Compute losses L_m and L_u based on S_m and S_u

 Compute total loss L as in Eq. (23)

 Update the network based on L

end

end

3.5. Loss function

The loss function consists of five parts: multimodal task loss L_m , unimodal task loss L_u , similarity loss L_s , diversity loss L_d and reconstruction loss L_r . These are combined as follows:

$$L = L_m + W_i^s L_u + \alpha L_s + \beta L_d + \gamma L_r \quad (23)$$

where the weights α , β and γ are used to adjust the contribution of each loss term to the total loss. $W_i^s = \tanh(|y_i^s - y_m|)$ represents the degree of difference between the labels of the unimodal task and the multimodal task: the greater the difference, the higher the weight.

3.5.1. Multimodal task loss

The multimodal task loss measures the performance of the model's multimodal predictions by calculating the absolute error between the predicted values and the true labels as follows:

$$L_m = \frac{1}{N} \sum_{i=1}^N |\hat{y}_m^i - S_m^i| \quad (24)$$

where \hat{y}_m^i is the multimodal prediction of the i th sample by the model as described in Algorithm 1, and S_m^i is the multimodal label of the i th sample.

3.5.2. Unimodal task loss

The unimodal task loss measures the performance of the model's predictions for each unimodal by calculating the absolute error between the predicted values and the true labels.

$$L_u = \sum_u |\hat{y}_s^i - S_u^i| \quad (25)$$

where \hat{y}_s^i is the unimodal prediction of the i th sample by the model, S_u^i is the unimodal label of the i th sample, and $u \in \{t, a, v\}$.

3.5.3. Similarity loss

The similarity loss is used to reduce the differences in the shared representations between each pair of modalities, ensuring that the modality-invariant representations of different modalities are as similar as possible, which aids in the fusion of invariant representations. The central moment discrepancy (CMD) is used as follows to measure the differences between the invariant representations of different modalities:

$$L_{sim} = \frac{1}{3} \sum_{(m1, m2)} CMD_K(h_c^{m1}, h_c^{m2}) \quad (26)$$

where h_c^m represents the modality-invariant representation of modality m , K is the order calculated in CMD, and $(m1, m2) \in \{(t, a), (t, v), (a, v)\}$.

3.5.4. Diversity loss

The diversity loss applies a relatively loose soft orthogonal constraint to ensure that two representations capture different aspects of the input data, thereby avoiding redundancy of information and ensuring that each representation captures unique information. The formula is as follows:

$$L_{diff} = \sum_m \|H_c^\top H_p\|_F^2 + \sum_{(m1, m2)} \|H_p^{m1\top} H_p^{m2}\|_F^2 \quad (27)$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm, H_c represents the feature representation matrix of the invariant modality, H_p represents the feature representation matrix of the specific modality, H_p^m represents the feature representation matrix of the specific modality m , and $(m1, m2) \in \{(t, a), (t, v), (a, v)\}$.

3.5.5. Reconstruction loss

Due to the orthogonal constraints between modalities, without additional measures, the encoder might generate orthogonal but unrepresentative vectors that meet mathematical orthogonality but do not contain practically useful information. Reconstruction loss is added to ensure that the modality-specific representations effectively capture the details of their corresponding modalities. The formula is as follows:

$$L_{recon} = \frac{1}{3} \sum_m \frac{\|u_m - \hat{u}_m\|_2^2}{d_h} \quad (28)$$

where $\|\cdot\|_2^2$ is the squared L^2 norm, u_m represents the original feature vector of modality m , and \hat{u}_m represents the reconstructed feature vector of modality m , $m \in \{t, a, v\}$.

4. Experiments

4.1. Datasets and benchmark models

4.1.1. Datasets

We use two classic datasets from the field of multimodal sentiment analysis, CMU-MOSI [30] and CMU-MOSEI [31], to verify the effectiveness of our model.

CMU-MOSI: The CMU-MOSI dataset is the first opinion-level annotated corpus for sentiment and subjectivity analysis in online videos. In addition to annotating subjectivity and sentiment intensity, it also includes visual features for each opinion and audio features annotated per millisecond. The CMU-MOSI dataset is sourced from YouTube movie review videos, containing a total of 93 randomly selected videos with 89 different speakers. The final CMU-MOSI dataset contains 3702 video segments, including 2199 opinion segments. Each sample's sentiment is annotated on a scale from -3 (highly negative) to 3 (highly positive).

CMU-MOSEI: The CMU-MOSEI dataset was created by collecting more utterances, samples, speakers and topics. It contains 23,453 speech videos from 1000 online YouTube speakers (57% male, 43% female). All sentences and utterances are randomly selected from different topics and individual videos. Following Ekman's theory of emotion, it includes sentiment annotations in the range of $[-3, 3]$ and emotion annotations in the range of $[0, 3]$, covering happiness, sadness, anger, disgust, surprise and fear.

4.1.2. Benchmark models

To demonstrate the effectiveness of our model in multimodal sentiment analysis tasks, we further compare it with the following baseline and state-of-the-art models:

TFN: Zadeh et al. [19] used a triple Cartesian product to capture the interactions between unimodal, bimodal and trimodal data, explicitly modeling the dynamic relationships within and between modalities. This is widely considered a classic work.

MFN: Zadeh et al. [32] used three independent LSTMs to model each modality separately, obtaining modality-specific interactions, and then used a memory attention network to capture the interactions between different modalities. Finally, temporal relationships were summarized through a multi-view gated memory.

RAVEN: Wang et al. [33] first analyzed fine-grained image and voice modalities accompanying text to model non-verbal representations. Then, based on non-verbal information, they altered word representations to capture the dynamic nature of non-verbal intent and obtained utterance representations.

Mult: Tsai et al. [34] used cross-modal Transformers to extract and fuse cross-modal feature information by using directed pairwise cross-modal attention to transform two modalities into the target modality. Then, the representations of each target modality were concatenated through a splicing operation to obtain the final multimodal representation.

MISA: Hazarika et al. [23] proposed a new framework (MISA), which projects each modality into modality-invariant and modality-specific subspaces. The former uses distribution similarity constraints to minimize heterogeneity gaps and learn their commonalities, while the latter is specific to each modality and learns private feature information of each modality.

ICCN: Sun et al. [35] proposed the Interaction Canonical Correlation Network (ICCN), which uses Deep Canonical Correlation Analysis (DCCA) to learn the relationships between text, audio and video for multimodal language analysis. Multimodal language analysis often considers the relationships between features based on text and those based on acoustical and visual properties.

Self-MM: Yu et al. [13] assigned additional modality-specific tasks for each modality and proposed generating unimodal labels in a self-supervised manner. By jointly training multimodal and unimodal tasks,

Table 1
Performance of all models on CMU-MOSI and CMU-MOSEI datasets.

Model	MOSI					MOSEI				
	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1
TFN	0.901	0.698	34.9	80.8	80.7	0.593	0.700	50.2	82.5	82.1
MFN	0.965	0.632	34.1	77.4	77.3	0.623	0.677	48.0	76.0	76.0
RAVEN	0.915	0.691	33.2	78.0	76.6	0.614	0.662	50.0	79.1	79.5
MuT	0.871	0.698	40.0	83.0	82.8	0.580	0.703	51.8	82.5	82.3
MISA	0.783	0.761	42.3	83.4	83.6	0.555	0.756	52.2	83.6	83.8
ICCN	0.862	0.714	39.01	83.07	83.02	0.565	0.713	51.58	84.18	84.15
Self-MM	0.708	0.796	46.67	85.46	85.43	0.531	0.765	53.87	85.15	84.90
ConFEDE	0.742	0.784	42.27	85.52	85.52	0.522	0.780	54.86	85.82	85.83
ULMD	0.700	0.799	47.81	85.82	85.71	0.531	0.770	53.81	85.95	85.91

Table 2
Module ablation experimental results.

Module	MAE	Corr	Acc-7	Acc-2	F1
ULMD	0.700	0.799	47.81	85.82	85.71
w/o separator	0.727	0.785	47.52	84.15	84.21
w/o interaction	0.737	0.783	44.9	83.99	83.96
w/o ULGM	0.759	0.779	43.73	83.54	83.54

the performance of multimodal sentiment analysis tasks was enhanced. This work is one of the main inspirations for the method in this paper.

ConFEDE: Yang et al. [24] proposed a unified learning framework for contrastive feature decomposition, which enhances the representation capabilities of multimodal information by jointly performing contrastive representation learning and contrastive feature decomposition. It decomposes the three modalities of video samples into similar and dissimilar features and conducts contrastive relationship learning centered on text.

4.2. Experimental settings and evaluation metrics

For training, the learning rate of the overall model is set to $1e-3$, and the learning rate of the BERT model is set to $5e-5$. Model parameters are optimized using Adam. The experiments are conducted on an Nvidia RTX 3090 GPU. Due to the different training and testing methods, multimodal sentiment analysis tasks can be divided into classification tasks and regression tasks. Therefore, in this paper, the experimental results of multimodal sentiment analysis tasks are evaluated in both classification and regression forms. For classification tasks, F1 score, binary classification accuracy (Acc-2), and seven-class accuracy (Acc-7) are used as evaluation metrics. For regression tasks, Pearson correlation coefficient (Corr) and mean absolute error (MAE) are chosen. Apart from MAE, the model's performance is proportional to the value of the evaluation metrics.

4.3. Results and analysis

To verify the performance of ULMD, comparative experiments were conducted on the CMU-MOSI and CMU-MOSEI datasets. The comparison results are shown in Table 1. Overall, in the comparative experiments on the CMU-MOSI dataset, the ULMD model achieved the best results. Specifically, in classification tasks, the ULMD model showed the greatest improvement, with an increase of 0.3% in Acc-2 and 0.19% in F1 scores compared to the best baseline model, ConFEDE. In multi-class classification (Acc-7), the ULMD model exhibited the most outstanding performance, with an improvement of 5.54%. On the CMU-MOSEI dataset, our method outperformed all other baseline models in both Acc-2 and F1 metrics. Additionally, our F1-score, Acc-7 and MAE metrics surpassed most baseline models.

Compared to the other two top-performing models, Self-MM and ConFEDE, the improvements of the ULMD model in various evaluation metrics are mainly attributed to the design of the multimodal fusion task and the modality representation separator. This also demonstrates

Table 3
Modality ablation experimental results.

Model	MAE	Corr	Acc-7	Acc-2	F1
M(TV+TA), T, A, V	0.700	0.799	47.81	85.82	85.71
M(TV+VA), T, A, V	0.716	0.794	44.61	85.06	84.96
M(TA+VA), T, A, V	0.732	0.793	44.75	84.91	84.89
M, T, V	0.723	0.789	47.38	84.3	84.28
M, T, A	0.737	0.792	43.15	83.84	83.9
M, V, A	0.774	0.789	42.71	84.45	84.52
M, T	0.708	0.795	45.63	85.52	85.37
M, A	0.705	0.797	46.36	85.82	85.70
M, V	0.727	0.795	45.77	85.06	85.01
M	0.742	0.792	44.46	83.84	83.83

the effectiveness of using multilayer perceptrons to separately fuse text and image, and text and audio to obtain sentiment polarity and sentiment intensity, as well as the ability of the modality representation separator to preserve modality consistency and differences.

4.4. Ablation experiments

4.4.1. Module ablation experiments

To test the overall architecture and the functionality of the components introduced in this paper, ablation experiments were conducted on the CMU-MOSI dataset. The specific results are shown in Table 2. We validated the effectiveness of the ULMD architecture through three experiments. Specifically, we (1) removed the modality representation separator and shared the underlying features in the form of hard parameter sharing; (2) removed the multimodal interaction module from the model, i.e., instead of exploring multimodal polarity and intensity through sentiment polarity and sentiment intensity vectors, we used simple concatenation of features; (3) removed the unimodal supervision label generation module from the model, i.e., unimodal training used only multimodal labels without the additionally generated unimodal labels.

From the results, it can be seen that the complete ULMD model achieved the best experimental results. Removing or replacing any module in the model leads to varying degrees of performance degradation, which validates the effectiveness of each module. The absence of the modality separator leads to a significant drop in model performance, indicating that the modality separator plays a key role in balancing modality consistency and differences. The lack of the interaction module shows that merely using simple concatenation of unimodal features makes it difficult to explore the complex interactions between modalities, thus failing to capture modality-invariant information. When the unimodal label generation module is removed, the performance of the model decreases significantly due to the lack of unimodal labels for the unimodal tasks.

4.4.2. Modality ablation experiments

To further investigate the role of each unimodal task and the multimodal interaction module, ablation studies were conducted on the CMU-MOSI dataset. The specific results are shown in Table 3, where

Table 4
Loss term ablation experimental results.

	MAE	Corr	Acc-7	Acc-2	F1
Full	0.700	0.799	47.81	85.82	85.71
w/o L_u	0.721	0.792	46.21	84.3	84.27
w/o L_s	0.715	0.790	44.46	85.37	85.34
w/o L_d	0.718	0.791	46.81	85.22	85.24
w/o L_r	0.757	0.782	44.17	83.69	83.77

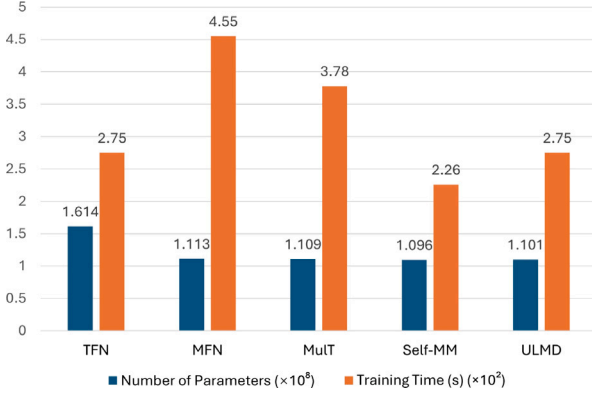


Fig. 4. Comparison of model complexity.

M, T, A and V represent the multimodal, text, audio and video tasks, respectively. M(TV+TA) indicates a multimodal task with text modality as the core, M(TV+VA) and M(TA+VA) follow the same logic, and M by default uses text modality as the core.

From the results, it can be seen that introducing unimodal subtasks can effectively improve model performance compared to single-task models, demonstrating the effectiveness of the multi-task learning framework. Additionally, among the text, image, and audio modalities, the text modality contributes the most to the model, which aligns with intuitive understanding and the design principle of using text as the core in our model.

4.4.3. Loss term ablation experiments

The impact of each loss term on the model's performance was evaluated through ablation experiments, with the results presented in Table 4. By setting the weights of different loss terms to zero, we effectively disabled them. The experimental results showed that the absence of unimodal label-based auxiliary learning significantly impacted the overall performance of the model. Compared to similarity loss and diversity loss, the model exhibited greater sensitivity to the absence of reconstruction loss. This could be due to the fact that, without the reconstruction constraint, while the model is able to learn the differences between various representations, these representations lack sufficient expressiveness.

4.5. In-depth analysis

4.5.1. Complexity analysis

To evaluate the complexity of the proposed method, we conducted a thorough analysis. The space complexity is measured by the number of parameters, while the time complexity is represented by the running time (with early stopping if no performance improvement is observed over several epochs). The test results on the CMU-MOSI dataset are shown in Fig. 4. Compared with other models, our method maintains a similar space complexity, while demonstrating better time complexity than most models, except for Self-MM. The slightly longer running time could be attributed to the additional computational demand required by ULMD for learning modality-invariant and modality-specific representations.

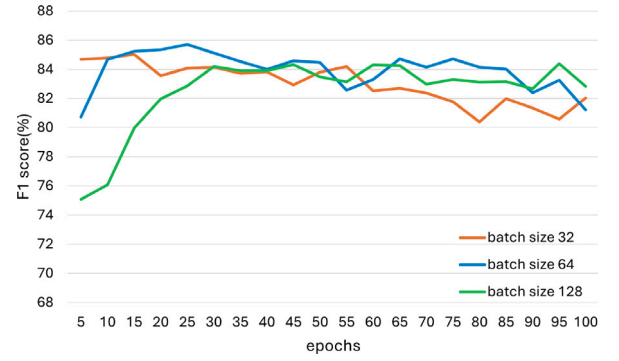


Fig. 5. Comparison of model performance with different batch sizes.

4.5.2. Hyperparameter sensitivity analysis

To examine the impact of key parameters on model performance, we compared the model performance corresponding to common batch sizes across different training epochs. The dataset used was CMU-MOSI, and F1 Score was adopted as the evaluation metric, as shown in Fig. 5. It can be observed that the model performs well overall across different batch sizes. However, when the batch size is set to 32, the model's performance appears to be more unstable. This instability may be attributed to the higher variance in gradient estimates for smaller batch sizes, while larger batch sizes can provide more stable gradient estimates.

4.6. Case study





To evaluate the quality of the unimodal labels, we selected several multimodal samples from the CMU-MOSI dataset and compared the unimodal labels generated during training with the human-annotated labels. The results are shown in Table 5.

In Case 1 and Case 2, both the unimodal and multimodal labels tend towards the same sentiment, indicating consistency between modalities. These two cases demonstrate that the unimodal labels generated by ULGM are valuable and generally consistent with the multimodal labels, positively contributing to task training. However, in Case 3, the multimodal label is -0.2 , but the text and audio modalities show positive emotions. Similarly, in Case 4, the multimodal label is 0.6 , while the image modality indicates negative emotions, showing inconsistency between modalities and a reverse bias towards the multimodal label. This suggests that the generated unimodal labels can adapt to special expressions such as sarcasm, providing more possibilities and variations for sentiment prediction, and helping to learn the characteristics of each modality.

5. Conclusion

This paper proposes a multimodal sentiment analysis method based on unimodal label generation and modality decomposition, trained using a multi-task learning framework. By introducing a modality representation separator and a unimodal label generation module, we divide modality representations into modality-invariant and modality-specific representations, which are then fed into multimodal tasks and unimodal subtasks, respectively, to assist in generating unimodal labels. This approach effectively alleviates issues of modality redundancy, modality heterogeneity, and the lack of unimodal labels in existing datasets. Comprehensive experimental evaluations conducted on the CMU-MOSI and CMU-MOSEI datasets validate the effectiveness of the proposed method.

Table 5
Case study analysis of real and predicted labels in MOSI dataset.

ID	Multimodal information	Video frame	Multimodal label	Text/Image/Audio labels
1	Text: It looks really good Image: Smiling Audio: Rising intonation		1.8	1.66/1.04/0.65
2	Text: Some really sucky acting Image: Frowning Audio: Downward intonation		-2.6	-1.98/-1.36/-0.95
3	Text: I mean it's pleasant to the story Image: Neutral expression Audio: Steady intonation		-0.2	0.33/-0.23/0.06
4	Text: And there were times that I thought it was funny Image: Downward gaze Audio: Neutral intonation		0.6	0.85/-0.12/0.18

CRediT authorship contribution statement

Linan Zhu: Supervision, Project administration, Conceptualization. **Hongyan Zhao:** Writing – original draft, Software, Methodology. **Zhechao Zhu:** Software, Conceptualization. **Chenwei Zhang:** Writing – review & editing. **Xiangjie Kong:** Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (No. 62176234, 62072409).

Data availability

The data used in this study are publicly available.

References

- [1] X. Wu, D. Hong, J. Chanussot, UIU-net: U-net in U-net for infrared small object detection, *IEEE Trans. Image Process.* 32 (2022) 364–376.
- [2] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, J. Chanussot, Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion, *IEEE Trans. Geosci. Remote Sens.* (2023).
- [3] X. Wu, D. Hong, J. Chanussot, Convolutional neural networks for multimodal remote sensing data classification, *IEEE Trans. Geosci. Remote Sens.* 60 (2021) 1–10.
- [4] S.K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, *ACM Comput. Surv. (CSUR)* 47 (3) (2015) 1–36.
- [5] J. Martinez-Miranda, A. Aldea, Emotions in human and artificial intelligence, *Comput. Hum. Behav.* 21 (2) (2005) 323–341.
- [6] Z. Wang, P. Wang, K. Liu, P. Wang, Y. Fu, C.T. Lu, C.C. Aggarwal, J. Pei, Y. Zhou, A comprehensive survey on data augmentation, 2024, *arXiv preprint arXiv:2405.09591*.
- [7] J. Deng, D. Hong, C. Li, J. Yao, Z. Yang, Z. Zhang, J. Chanussot, RustQNet: Multimodal deep learning for quantitative inversion of wheat stripe rust disease index, *Comput. Electron. Agric.* 225 (2024) 109245.
- [8] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, X.X. Zhu, Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks, *Remote Sens. Environ.* 299 (2023) 113856.
- [9] V. Pérez-Rosas, R. Mihalcea, L.P. Morency, Utterance-level multimodal sentiment analysis, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 973–982.
- [10] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: *2016 IEEE 16th International Conference on Data Mining, ICDM, IEEE, 2016*, pp. 439–448.
- [11] E. Shutova, D. Kiela, J. Maillard, Black holes and white rabbits: Metaphor identification with visual features, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 160–170.
- [12] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, *Multimed. Syst.* 16 (2010) 345–379.
- [13] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, (12) 2021, pp. 10790–10797.
- [14] K. Liu, X. Zhao, Y. Hu, Y. Fu, Modeling the effects of individual and group heterogeneity on multi-aspect rating behavior, *Front. Data Comput.* 2 (2020) 59–77.
- [15] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, *Inf. Fusion* 95 (2023) 306–325.
- [16] L. Zhu, M. Xu, Y. Bao, Y. Xu, X. Kong, Deep learning for aspect-based sentiment analysis: a review, *PeerJ Comput. Sci.* 8 (2022) e1044.
- [17] L.P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: *Proceedings of the 13th International Conference on Multimodal Interfaces*, 2011, pp. 169–176.
- [18] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, L.P. Morency, Deep multimodal fusion for persuasiveness prediction, in: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 284–288.
- [19] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.P. Morency, Tensor fusion network for multimodal sentiment analysis, 2017, *arXiv preprint arXiv:1707.07250*.
- [20] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, 2018, *arXiv preprint arXiv:1806.00064*.
- [21] L. Zhu, M. Xu, Y. Xu, Z. Zhu, Y. Zhao, X. Kong, A multi-attribute decision making approach based on information extraction for real estate buyer profiling, *World Wide Web* 26 (1) (2023) 187–205.
- [22] T. Baltrušaitis, C. Ahuja, L.P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (2) (2018) 423–443.
- [23] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.
- [24] J. Yang, Y. Yu, D. Niu, W. Guo, Y. Xu, Confede: Contrastive feature decomposition for multimodal sentiment analysis, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 7617–7630.
- [25] Y. Hwang, J.H. Kim, Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis, in: *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 35–46.
- [26] M. Li, Z. Zhu, K. Li, L. Zhou, Z. Zhao, H. Pei, Joint training strategy of unimodal and multimodal for multimodal sentiment analysis, *Image Vis. Comput.* 149 (2024) 105172.
- [27] Z. Li, Q. Guo, Y. Pan, W. Ding, J. Yu, Y. Zhang, W. Liu, H. Chen, H. Wang, Y. Xie, Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis, *Inf. Fusion* 99 (2023) 101891.
- [28] M. Li, D. Yang, X. Zhao, S. Wang, Y. Wang, K. Yang, M. Sun, D. Kou, Z. Qian, L. Zhang, Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12458–12468.
- [29] J. Hou, N. Omar, S. Tiun, S. Saad, Q. He, TCHFN: Multimodal sentiment analysis based on text-centric hierarchical fusion network, *Knowl.-Based Syst.* 300 (2024) 112220.
- [30] A. Zadeh, R. Zellers, E. Pincus, L.P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, *arXiv preprint arXiv:1606.06259*.

- [31] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [32] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.P. Morency, Memory fusion network for multi-view sequential learning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, (1) 2018.
- [33] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, (01) 2019, pp. 7216–7223.
- [34] Y.H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.P. Morency, R. Salakhutdinov, Multi-modal transformer for unaligned multimodal language sequences, in: *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2019, NIH Public Access, 2019, p. 6558.
- [35] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, (05) 2020, pp. 8992–8999.